MODELING THE IMPACT OF OPEN SOURCE SOFTWARE: A STUDY OF R PACKAGES

Ronnie Fecso (VT), Claire Kelling (PSU) with Gizem Korkmaz and Stephanie Shipp (SDAL)
Sponsor: Carol Robbins, The National Center for Science & Engineering Statistics at NSF

SDAL SOCIAL & DECISION ANALYTICS LABORATORY
DATA SCIENCE FOR THE PUBLIC GOOD

Research Question

This project aims to identify the factors that affect the impact of Open Source Software (OSS), measured by the number of downloads and citations, with a case study of R packages. We generate the dependency network of the packages collected from Depsy.org, and develop statistical models that use the network characteristics and the package attributes.

Background: The programming language R was created in 2000 by two professors at the University of Auckland for their statistics students. It is one of the fastest growing programming languages due to low barriers to entry. It is widely used for data analysis, visualization, and statistical modeling.

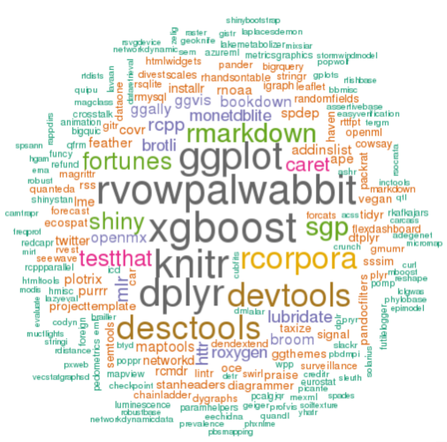
Data Source: Depsy.org

Depsy.org, initially NSF-funded, is a website that compiles R and Python packages to quantify coding impact in the scientific community. It includes detailed information such as contributors, commits, downloads, citations, and stars (identifying active development).

Data Collection Process:

- Gathered all of the R packages listed on CRAN (10,926 packages listed as of July 11th, 2017)
- Scraped the characteristics from the JSON page affiliated with each R package from Depsy
  - 9,810 of these packages (90%) were on Depsy (last update in Sept. 2015) with 24,000+ affiliated contributors

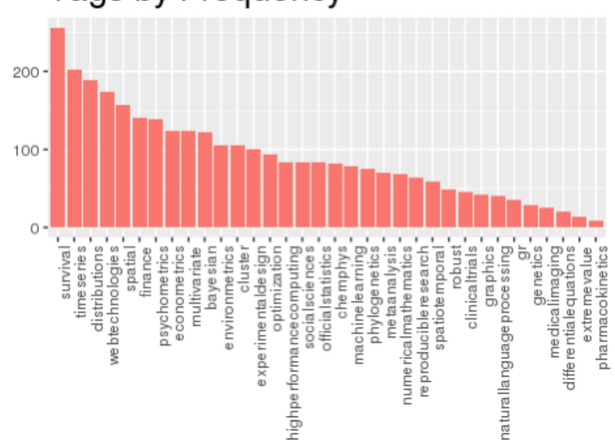
R Packages by #contributors



Tags on Depsy.org

- There are 34 tags on Depsy
- Average number of tags per package: 1.3
- Average number of packages per tag: 90.8
- 2,429 packages (25%) have tags

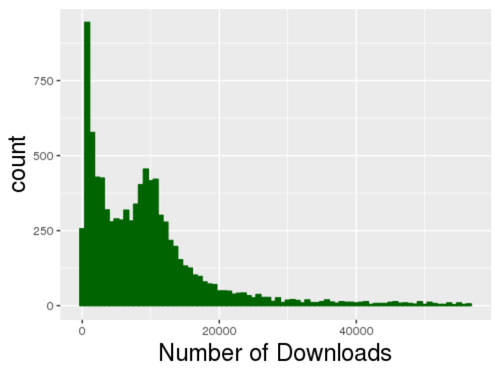
Tags by Frequency



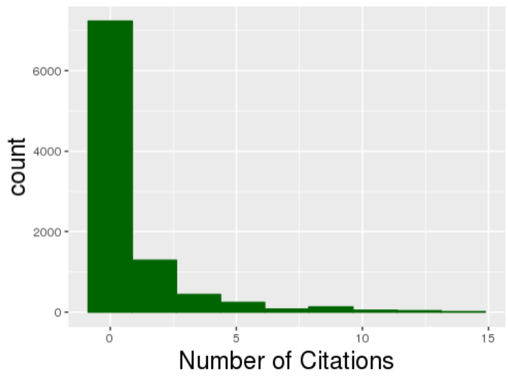
Measuring Impact: Downloads and Citations

Top Downloaded Packages	Number of Downloads
Rcpp	6,683,565
ggplot2	6,255,500
stringr	5,366,703
plyr	5,345,308
digest	5,251,824

Top Cited Packages	Number of Citations
vegan	4,275
lme4	4,023
nlme	2,916
ggplot2	1,702
gplots	1,307



The distribution of the number of downloads is skewed and bimodal. The average number of downloads is ~58K, and the average number of citations is ~6.84 per package.

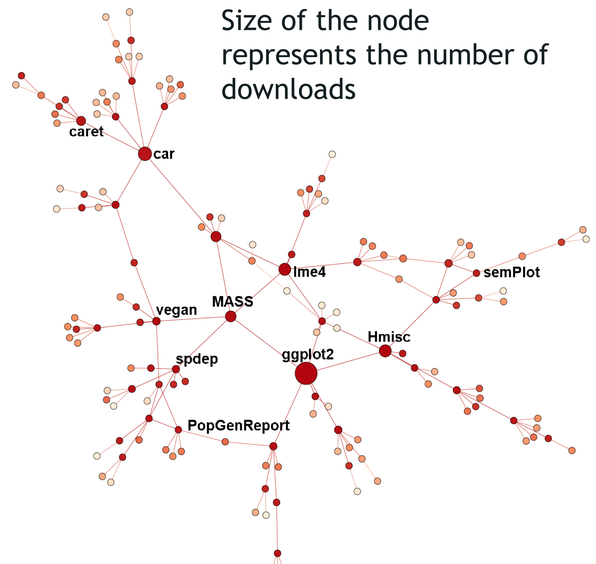


The Network of R Packages

We generate the dependency network of R packages where a directed edge $i \rightarrow j$ indicates that the package j requires i to be installed to function. The figure below uses a subgraph (starting at MASS as the root node) with 149 nodes for illustrative purposes.

Full Network Characteristics:

- 7,389 nodes
- 20,235 directed edges
- Average indegree = outdegree = 2.74
- 56 weakly connected components
  - largest component has 7,261 packages
  - Remaining components include 2 to 4 packages (mean 2.33)



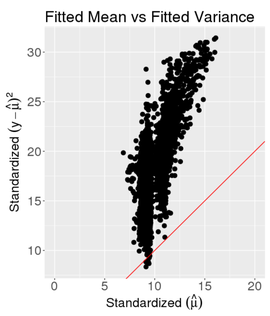
Size of the node represents the number of downloads

The Model & Findings

We develop two Quasi-Poisson models[1] with the number of downloads (Model 1) and the number of citations (Model 2) as the dependent variables, y. We let E(y)=μ and Var(y)= $\theta\mu$. We assume that $y_i \sim Poisson(\mu_i, \theta)$ and let the mean μ_i for the ith observation vary as a function of the p covariates as follows:

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i})$$

where the network characteristics and the package attributes are used as covariates.



The findings are summarized in the table below.

Parameters	Model I: #Download	Model II: #Citation	Parameters	Model I: #Download	Model II: #Citation
Intercept	9.04***	1.03***			
Network Variables					
Outdegree	9.32 e-3***	0.01**	Clustering	-1.34***	-3.41***
Indegree	-	-	Authority	-	56.9***
Closeness	0.91***	1.84***	Hub	-15.51***	-13.9***
Betweenness	-	1.22 e-4***	Component ID	-4.12 e-2***	-
Eigencentrality	0.69*	-	Modularity Class	-	-0.03***
Page Rank	-	-4.76 e+3**	Eccentricity	0.69***	0.59***
Strong Component ID	6.22 e-05***				
R-Package Variables					
# Authors	-1.44 e-02***	-	# Stars	-	-2.58 e-3***
# Commits	-5.53 e-4***	1.01 e-4***	# Citations	-1.69e-07*	-
# Committers	-	-	# Downloads	-	2.25 e-4***
# Contributors	4.06 e-2***	2.56 e-2***			

Significance codes: 0 (***), 0.001 (**), 0.01 (**), not significant (-)

We find that the network centrality of a package (the number of packages that depend on it, and how they are connected) as well as the number of contributors and commits are important factors in showing the impact of open source software, measured by the number of downloads and citations.

Definitions: Outdegree (indegree, resp.) is the total number of outgoing (incoming) links. Closeness centrality measures how close a node is to every other node. Betweenness is a measure of being connected to other nodes that are not connected to each other (as a bridge). Eigencentrality of a node takes into account the centrality of its neighbors. PageRank depends on (i) the number of links the node receives, (ii) the number of links given out by its neighbors, (iii) the centrality of its neighbors. Clustering coefficient quantifies the degree to which a node's neighbors are connected. Authority measures the value of information stored at the node; hub measures the quality of its links. Eccentricity captures the node's distance from the furthest node. Modularity is how the network decomposes into subnetworks.

Future Work

- Develop methods to measure the cost and the economic impact of OSS projects.
- Apply similar methods to measure the impact of Python packages.

[1] Ver Hoef, J. M. and Boveng, P. L. (2007), Quasi-Poisson vs Negative Binomial Regression: How Should We Model Overdispersed Count Data?. Ecology, 88: 2766–2772. doi:10.1890/07-0043.1